Utrecht University

Master Applied Data Science

July 2023

Evaluating the Effectiveness of the Topic Models LDA and BTM for Uncovering Topics in Open-Ended Employee Engagement Survey Responses

Carlijn Diederiks

5431379

Supervisor: G.A.W. Vreeswijk

2nd Examiner: D.P. Nguyen

Table of Contents

1. Introduction

The productivity of organizations is strongly determined by the efforts and engagement of its employees (Musgrove, Ellinger, & Ellinger, 2014). Engaged employees show an emotional attachment to their organization and are willing to go above and beyond their job responsibilities. As a result, organizations with engaged employees have a decreased turnover rate, high productivity and increased profitability (Markos & Sridevi, 2010). Welch (2011) describes engagement as a dynamic, changeable psychological state that connects employees with their organizations. Due to the dynamic nature of employee engagement, it is important to measure this frequently. A way to do this is with an employee engagement monitor, which is used to measure, understand and improve the level of engagement among employees. This is often measured with a survey with both open-ended and closed-ended questions, about for example job satisfaction and motivation.

Open-ended questions allow respondents to provide their responses in their own words without any restriction. The answers can vary in length, content and format. The reason why open-ended questions are included in survey instruments, is to capture information that is not represented in closed-ended questions or could not be foreseen by the designer of the survey. Another reason to include those questions is that it enables respondents to offer explanations on their chosen option from the closed-ended questions or give some additional information. The use of open-ended questions could even give a more human face to the survey instrument (Fielding, Fielding & Hughes, 2013). However, analyzing these types of questions can be challenging due to the lack of structure. Miles (1979) described the analysis of open-ended questions as a “nuisance”, because the analysis of qualitative data is labor-intensive and there is not a standard analytical method to analyze those questions.

The aim of this paper is to identify topics in a large dataset containing both short and long text answers to open questions of an engagement monitor developed by the company Scorius. Knowledge of Natural Language Processing (NLP) can offer a solution to automate this analysis that was previously performed manually. NLP is an evolving research field that combines the knowledge of linguistics, computer science and artificial intelligence, to obtain meaning from natural human language with machines. NLP has many applications that could be used to analyze text data, for example sentiment analysis, language detection and topic modeling. Although there is not an established analytical method to analyze qualitative survey data yet, topic modeling is a powerful tool for text mining and can be used to discover topics from a collection of text documents (Jelodar et al., 2019).

The most popular and highly studied model in the topical modeling field is the Latent Dirichlet Allocation (LDA) model (Blei, Ng & Jordan, 2003). This model has the capability to generate a set of topics, which are understandable at individual level and describes the entire collection of text documents. Additionally, it can handle large collections of data and does not require any labeled input. LDA has been used in the analysis of qualitative data, for example in the research of Nanda et al. (2021). The model was used to identify topics in open-ended questions about an online course. LDA succeeded in classifying the common themes and topic modeling serves as a good starting point. However, this research concluded with the statement that to derive meaningful conclusions, human interpretation is necessary.

Yan, Guo, Lan and Cheng proposed in 2013 a method as an alternative to LDA. This model is called biterm topic model (BTM) and is created to work with short texts. This model uses word embeddings to advance topic modeling. In the research of Vidal, Ares and Jaeger (2022), the capabilities of BTM in analyzing responses to open-ended survey questions were explored in open-ended questions about vertical farming. Based on their results, the potential for revealing underlying topics is promising. The BTM model predicted consistent, coherent topics and revealed the aspects that were discussed in the responses.

The potential of both the LDA model and the BTM model to offer valuable insights into the underlying topics of open-ended questions is promising. However, previous research has primarily focused on datasets comprising short texts and has not explored the analysis of combined short and long texts. Exploring methods that can effectively handle both types of texts can enhance the understanding of the complexities present in textual data of varying lengths. Additionally, the use of these models in analyzing employee engagement has not been explored thus far, presenting a promising opportunity for research and practical implementation. This leads to the following research question: "Are the topic modeling methods LDA and BTM effective to use to gain insight in the topics present in unstructured answers to open ended questions in employee engagement surveys?"

This research is conducted on behalf of the company Scorius and offers a way to improve the analysis of text data obtained by their engagement monitor. The company Scorius combines their knowledge of IT, data and organizational psychology to support large organizations with collecting employee data and with the analysis of this data. It aims to provide a topic list which could be used to extract useful information from such questions, which Scorius can utilize for qualitative analysis of their engagement monitor.

The next section of this paper is the theoretical framework, exploring the origins of the two models and providing a clarification of the underlying principles. Section 3 describes the data used in this research. In section 4, the steps in preprocessing the data are explained, in order to use it as input for topic modeling. The method to tune the models is explained and the method for the evaluation of the topic modeling methods is described as well. The results are presented in section 5. Finally, the discussion in section 6 interprets the results and gives an answer to the research question. Limitations, further research, ethical considerations and a conclusion are also included in this section. The appendix contains a link to Github, where the code used for this research can be downloaded.

2. Theoretical framework

The LDA model and the BTM model are topic modeling methods. This section explains the origins of the models and provide a comprehensive explanation of their underlying principles and operational mechanisms.

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a Bayesian topic modeling method, which is used for discovering topics in a set of text documents (Blei et al., 2003). The word "latent" in the name refers to the fact that the topics are not known beforehand and cannot be measured directly. Therefore, this method is an unsupervised learning technique, that uses a statistical model based on the Dirichlet probability distribution.

Many models have preceded LDA in the field of information retrieval, such as Tf-idf reduction, LSI and pLSI. However, each model fails to offer a probabilistic model at the level of documents (Blei et al., 2003). The fundamental idea of the LDA model is that it views each document as a combination of underlying topics. These topics are comprised of words that commonly appear together in the dataset and are utilized by the model to learn. Through iterative analysis of word co-occurrence in documents, the LDA model aims to identify these topics and represent each document as a mixture of various topics, each with corresponding weights.

The LDA model assumes that documents consist of "bag of words", which means that the order of the words is not relevant. A second assumption is that the order of the documents does not matter as well. Lastly, it is assumed that the number of topics is known and fixed (Blei, 2012).

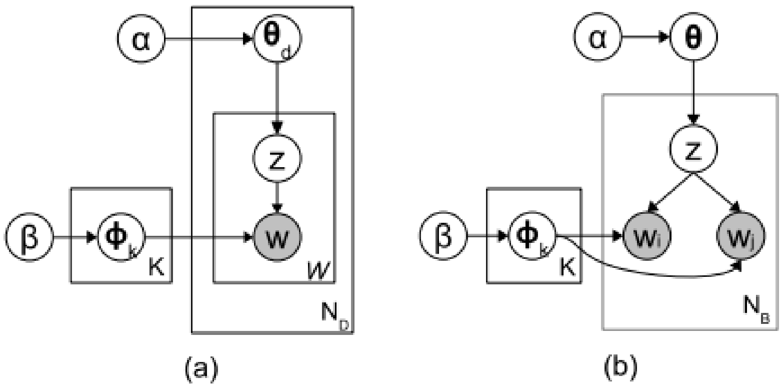

Fig 1. *Graphical representation of (a) the LDA model and the (b) BTM model.*
Note. Adapted from Cheng et al., 2014.

The left graph in Figure 1 shows graphically how the LDA algorithm works. The shading of the W means that this variable, the words, is the only observed variable. The other variables are:

- Z is a topic indicator variable
- The document-specific topic distribution $\Theta_d \sim \text{Dirichlet}(\alpha)$
- The word distribution $\Phi_k \sim \text{Dirichlet}(\beta)$

The parameters that have to be set in this model are α and β, which are the document-topic density and the topic-word density.

The three rectangles in the model denote replication of the model within it. So, the number of replications are given by:

- K is the number of topics chosen
- N_D is the number of documents
- W is the number of single words

The iterative process that is shown in Figure 1 starts with looping through each document and randomly assign each word to one of the K topics. Then looping through each word again, unassign the topic and reassign the topic based on the conditional probability that the word takes on each topic. To get this probability, two proportions had to be computed. $p(w|t_k)$ is the proportion assignments over all documents to topic t_k for this word w. The second one, $p(t_k|d_i)$, is the proportion of words in document d_i that are assigned to topic t_k. This conditional probability is based on multiplying $p(w|t_k)$ and $p(t_k|d_i)$. The word will be reassigned to the topic with the highest conditional probability. When this is done for all words, this loop starts over again as many times as the number of iterations. Since the first step was randomly creating topics, the generated topics at the start were not coherent. But with the iterations, the words gravitate towards each other and topics become more dispersed.

2.2 Biterm Topic Model

The biterm topic model (BTM), proposed by Yan, Guo, Lan and Cheng (2013), is the first principled approach for topic modeling with a focus on general-domain short texts. Even though LDA is a conventional method, it might not work well with short text, such as tweets. The reason for this is the sparse word co-occurrence in each short document. BTM is a method that models the biterms (i.e., unordered word pairs that co-occurred) in a set of text documents. In contrast with the LDA model, the BTM has been developed to discover the patterns in the whole set, instead of focusing on the patterns in the separate documents. That makes BTM suitable to learn topics over shorts text documents. Furthermore, BTM is characterized by its simplicity and ease in implementation. As with LDA, a Dirichlet word probability distribution is assumed.

The trainings data for the BTM model is created by extracting any two distinct words in a document as a biterm. For example, in the short sentence: "IT resources available.", three biterms are extracted, namely "IT resources", "IT available" and "resources available". In this model, it is assumed that these biterms are generated independently. However, two biterms in the same document are strictly speaking not independent, but this is assumed for simplicity of the model. The three assumptions of the LDA model are also applicable here.

As can be seen in Figure 1, the model of BTM and LDA look similar, nevertheless there are some differences. N_B refers to the number of biterms and instead of looking at single words, the biterms are used as input variable. Another major difference is that the BTM model considers the whole set of documents as a mixture of topics, instead of each document. Figure 1 visualizes this contrast. In the LDA model, the topic indicator Z of word W depends on Θ_d, which means that this relies on other words in the same document. While in the BTM model, Z is sampled from a global topic distribution θ (Yan et al., 2013). This has also implications on computing the conditional probabilities needed for learning the topics. Because we cannot get the topic proportions of documents, the empirical distribution of biterms in document d_i is taken as an estimation: $p(b|d_i)$. The conditional probability is based on multiplying $p(t_k|b)$ and $p(b|d_i)$, with biterm b and topic t_k.

3. Data

This research is conducted in cooperation with the company Scorius, that has provided a dataset with raw text data. The data was generated by open-ended questions asked to employees in multiple big companies in the interest of monitoring employee engagement. These questions were asked to a sample of employees within these companies. The answers to these questions were stored in a database and accompanied by employees' demographics and functions within the company. Due to privacy measures, the data used for this research only contains two columns, the questions and the answers. Additionally, Scorius manually removed parts of the answers that contained names or other information that refers to the identity of the employee. In the dataset, anonymized words were indicated by the reason for anonymizing (e.g., company name) and asterisks (*) surrounding them. The dataset included 11.403 rows and each row contains the answer from a unique employee. In topic modeling, the concept "document" is used to describe text fragments that are used as input for the model. In this dataset each row of the dataset is considered as a "document". Table 1 gives an idea of what the dataset looks like.

Table 1

A snippet of the provided dataset, with two sample questions and three sample answers.

question_text	*answer*
Wat kan *Bedrijf* doen om jouw betrokkenheid te vergroten?	Medewerkers op de werkvloer betrekken bij het nemen van beslissingen, luisteren naar argumenten die van invloed zijn op de te nemen beslissingen.
Wat kan *Bedrijf* doen om jouw betrokkenheid te vergroten?	Improve on local presence Equality for women
Wat kan *Bedrijf* doen om jouw betrokkenheid te vergroten?	Ik krijg binnen *Bedrijf* de mogelijkheid om te ontplooien, echter er wordt verder niets mee gedaan. Ik krijg niet de mogelijkheid om door te groeien en dat is frustrerend. De frustratie van het niet verder kunnen groeien is de reden dat ik (op dit moment) niet volmondig kan zeggen dat ik trots ben bij *Bedrijf* te werken.
Heb je opmerkingen, tips en/of aanvullingen? Geef dit dan hieronder weer:	Dat de collega.s op tijd komen op hun werk.
Heb je opmerkingen, tips en/of aanvullingen? Geef dit dan hieronder weer:	Het maken van echte keuzes in de organisatie is lastig evenals het elkaar aanspreken op zaken die niet goed gaan om zo de kwaliteit, efficiency en samenwerking te verbeteren.
Heb je opmerkingen, tips en/of aanvullingen? Geef dit dan hieronder weer:	n.v.t.

The dataset contains 32 different questions (see appendix A), some about specific subjects, for example "What can the company do to support you even more with the theme of vitality?". Others are more generic, such as "Do you have any comments, tips and/or additions?". The employees did not get any restrictions for answering those questions, which results in answers that differ in length, contents and text structure.

The answers have character lengths in a range from 1 to 2810 characters, the distribution is shown in Figure 2. The average character length is 159 characters, and the median is 86 characters. This indicates that there are more answers shorter than the mean character length.

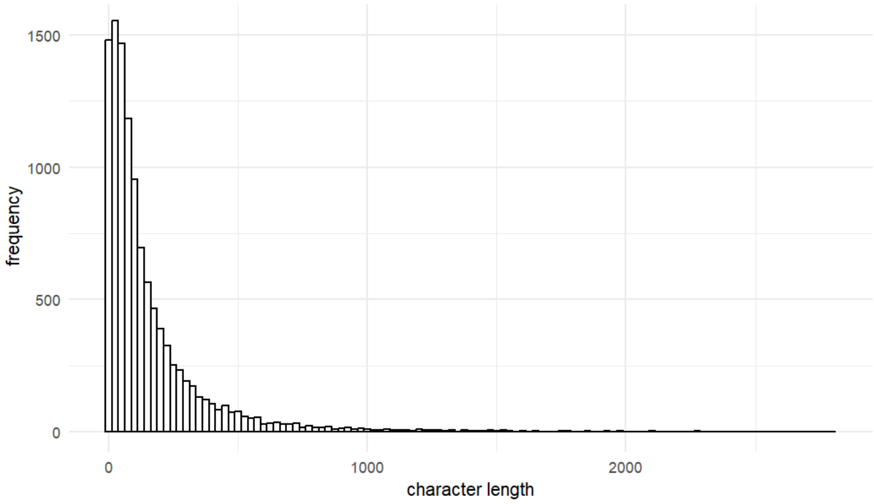

Fig 2. *Character distribution.*

The answer column involves mainly Dutch text answers, but for example also German, Turkish and English answers. Besides that, there is considerable amount of text that implies that the person does not have an answer on the question, as given by answers such as “no clue” or “not applicable”.

4. Method

In this section, the preprocessing phase is explained, which is required before using the data as input for topic modeling. Furthermore, the method to tune and evaluate the models is described and the data science problem proposed by Scorius is explained. All analyses were performed using R Statistical Software and Python (see appendix B).

4.1 Preprocessing the data

For applying the unsupervised learning method topic modeling, input with a good quality is needed. To achieve this, a preprocessing phase is needed to put the raw data in a tidy format.

First, the missing data is handled. 133 rows were deleted due to the fact that no answer has been given to the question and this will not give any valuable information in the topic modelling context. After this, the next steps are taken to clean the answer column:

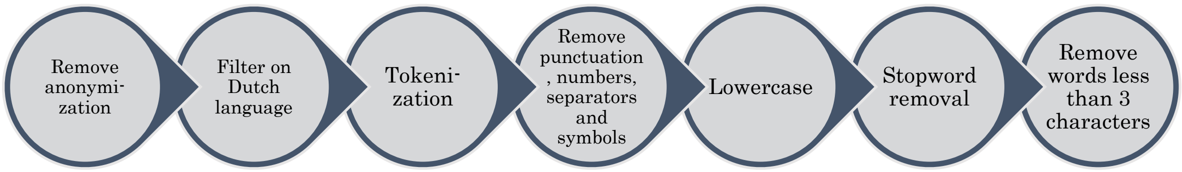

Fig 3. *The steps in the preprocessing phase*

Remove anonymization: The words that are anonymized by Scorius are removed, such as *company*, because they do not contain any useful information anymore after being anonymized.

Filter on Dutch language: The answers were given in different languages, but for better performance of the topic models, only the Dutch sentences are used. For filtering the sentences in Dutch, a Naïve Bayesian classifier is used. This method scored for each sentence the 3-6 most likely languages and their log probabilities. The language with the highest log probability is used to classify the language of the sentence. Assessed through eyeballing, this method classifies accurately. 16% of the sentences were not classified with a language, because the language could not be identified. These were mainly short sentences of one or two words. 48% of the dataset is excluded due to being not classified or classified as non-Dutch sentences.

Tokenization: One of the assumptions of both the LDA and the BTM model is that the words in the documents are a bag of words. With tokenization the sentences are broken into individual words, while the order is not relevant anymore.

Remove punctuation, numbers, separators and symbols: These characters are removed, because they do not carry semantic information and hinder the analysis.

Lowercase: All uppercase characters are changed to characters in lowercase, to normalize the text input.

Stopwords removal: Stopwords are common words that are often excluded for text analysis as they carry minimal information and are generally not essential for understanding the meaning of a sentence. One of the goals of removing these words is to enhance the speed of execution by reducing the vector space. It also contributes to overall performance improvement (Kaur & Buttar, 2018). The stopwords are removed with the Dutch iso stopwords list, which is the most comprehensive collection of stopwords. Other words are added manually ("nvt", "idee", "weet", "soms", "ten", "qua", "één"), while these words do not contain any useful information in the context of our data. Such as “nvt”, which is the Dutch abbreviation of does not apply.

Remove words less than 3 characters: Words with one or two characters are removed, because they carry limited semantic meaning and help reduce noise in the data.

Another preprocessing method which is frequently used in topic modeling is stemming which reduces words to their root form. But according to the research of Schofield and Mimno (2016), stemming has no meaningful improvement in topic coherence and can even degrade the topic stability. With this knowledge, stemming is not used in this research.

After the execution of the preprocessing steps, a document-term matrix is created. This is needed as input for the LDA model. The matrix captures the frequency of each term within a document, as can be seen in Table 2. In this step, terms are deleted that appear in less than 0.1% of the sentences. These infrequent words are removed to reduce the dimensionality of the matrix and to improve the quality of the input. These uncommon terms are often considered as noisy and irrelevant. In order to maintain consistency in the input for both the LDA and BTM models, these filtered words are also removed from the tokenized dataset, which serves as the input for the BTM model.

The documents which do not contain any words after the preprocessing, are removed. The cleaned dataset consists of 5702 documents and 1765 unique words.

Table 2

The document term matrix

	Features					
Docs	afspraken	nakomen	terugkoppeling	geven	vragen	…
text1	1	1	1	1	1	
text2	0	0	0	0	0	
text3	0	0	0	0	0	
text4	0	0	0	0	0	
text5	0	0	1	0	0	
text6	0	0	0	0	0	
…						

4.2 Tuning the parameters

One of the assumptions of both the LDA and the BTM model is that the number of topics k is known. However, this is not a predefined number and an optimal number had to be chosen before running the models. The number of topics determines how specific discovered topics are. A small number of topics often leads to very general topics (Yan et al., 2013). Determining this number of topics is considered a difficult task and many alternative approaches have been proposed. The best approach has not been found yet and could also depend on the nature of the text (Bystrov et al., 2022).

To find the optimal number of topics for the LDA model, a combination of three theoretically grounded approaches has been used. Firstly, the log-likelihood of the data was calculated for different number of topics. It evaluates how well a topic model represents the given set of documents. A higher value on log-likelihood predicts a better fit (Griffiths & Steyvers, 2004). Secondly, the optimal number of topics can be based on topic density. Small inter-cluster density predicts stability of the topics, so this score has to be minimalized (Cao et al., 2009). Lastly, divergence score can be used to find the right k. Arun et al (2010) observed that when using a non-optimal number of topics, the divergence values tend to be higher. So, minimizing the divergence score is desirable. By comparing these three metrics and finding the extremum, the optimal value for k can be approximated.

For the BTM model, the Rényi entropy score is used, which has been researched by Koltcov (2018) as a way to approximate k. This score is an index of diversity and low entropy values indicate that the assigned topics are more focused and have less overlap.

With the chosen k, the document-topic density α and the topic-word density β had to be tuned. A high value on α places more weight on having each document composed of many dominant topics. A high value on β places more weight on having each topic composed of most of the words. A tuning process grid search is used, which means that combinations of multiple values of α and β are compared on their performances. Values of α are chosen in a range from [0.01, 0.05, 0.1, 0.5, 50/k], with the inclusion of 50/k implemented based on the recommendation from previous research findings (Griffiths & Steyvers, 2004). Values of β are chosen in a range from [0.01, 0.05, 0.1, 0.5]. For each combination of α and β the coherence score is calculated. This automated metric for topic quality evaluation is proposed by Mimno et al (2011). It is based on the idea that words associated to a single concept are likely to appear together in the same documents. Topic coherence measures for each topic separately the degree of semantic similarity between the top words in the topic and is defined with this formula:

$$C\left(z;\ V^{(z)}\right) = \sum_{t=2}^{T} \sum_{t=1}^{t} \log \frac{D(v_m^{(z)}, v_l^{(z)}) + 1}{D(v_l^{(z)})},$$

given a topic z and its top T words $V^{(z)} = (v_1^{(z)}, \ldots, v_T^{(z)})$. $D(v)$ is the document frequency of word v. The mean of the coherence scores per topic is calculated and used for comparison.

Due to the random initialization of topic assignments, each run of the topic modeling process results in a different outcome. As a result, the coherence scores also vary for each model. To take an informed and robust decision regarding the choice of parameters, the average of multiple models per parameter combination are calculated. For LDA, this is an average of ten different models, and for BTM, it involves five different models, as the running time is considerably higher. The highest average topic coherence score is used for choosing the best performing parameters α and β.

4.3 Evaluation

When having the parameters set, the two proposed models are run. The Gibbs sampling algorithm is used for iteratively updating the topic assignments. 1000 iterations are chosen to let the words gravitate towards each other and create dispersed topics. A seed was set to make the outcome of the models reproducible. To evaluate which model performs the best in modeling the topics in the dataset, quantitative and qualitative measures can be used.

The coherence score as described above, can also be used to compare the topic quality of the LDA and BTM model in a quantitative way. The highest average topic score could be an indicator for the best model. Research has shown that the coherence score is highly correlated with human judgement of the coherence of the topics (Yan et al., 2013). Three coherences scores per model are computed, based on the top five, ten or twenty topic words. These top words are ranked on conditional probabilities. This is the probability of a word occurring within a particular topic, given the topic assignment for the other words in the document. The words with the highest probabilities are used to compute the coherence score. The coherence score is measured per topic, but the mean of those and their corresponding 95% confidence intervals are used as a metric to compare the models. Another quantitative measure that could be used is the perplexity score. However, the decision has been made not to utilize this metric due to its potential to contradict human interpretation (Newman et al., 2010).

Another way to evaluate the topics is to label the topics generated by the two methods. A domain expert, an employee of the company that provided the data, is asked to label the topics using the top ten corresponding topic keywords. When the expert determines there is no label that fits the words, no label is assigned, which suggests that the topic is challenging to interpret. In addition, for each labeled topic, he is asked to indicate whether he is satisfied with the given label or not. This determines whether the label is a good fit to the words.

Lastly, the technique of visually inspecting the top ten words associated with each topic, commonly known as eyeballing, is performed. This qualitative approach allows for the assessment of the uniqueness and specificity of the topics.

4.4 Data science problem

The data science problem of this research is to get insights into the topics presented in unstructured answers to open ended questions in employee engagement surveys. The best evaluated topic model method is used to create a list with the labeled topics and the twenty corresponding top words. Assessed is whether this generated list is good enough to classify the answers with corresponding topics or that the list needs some manual finetuning. This list can be used by Scorius as input for analyzing the employee engagement data.

5. Results

The results of determining the optimal number of topics and the tuning of the parameters are presented in this section. With these optimized values, the LDA and BTM model are run and evaluated with the described evaluation metrics.

5.1 Tuning the parameters

To determine the optimal number of topics *k* for the LDA model, a combination of three metrics is used. Figure 4 shows the values for different number of topics. The Arun2010 divergence metrics had to be minimized and the graph exhibited an initial sharp decline from the beginning and is followed by a flattening starting around 25 topics. The CaoJuan2009 density demonstrated a decrease in values when the number of topics was in the range of 15 to 25 topics. The Griffiths2004 log-likelihood metrics was maximized at 25 and at 40 topics and after 40 topics it showed a decline. By combining these three metrics, an approximation of the optimal number of 25 topics can be obtained.

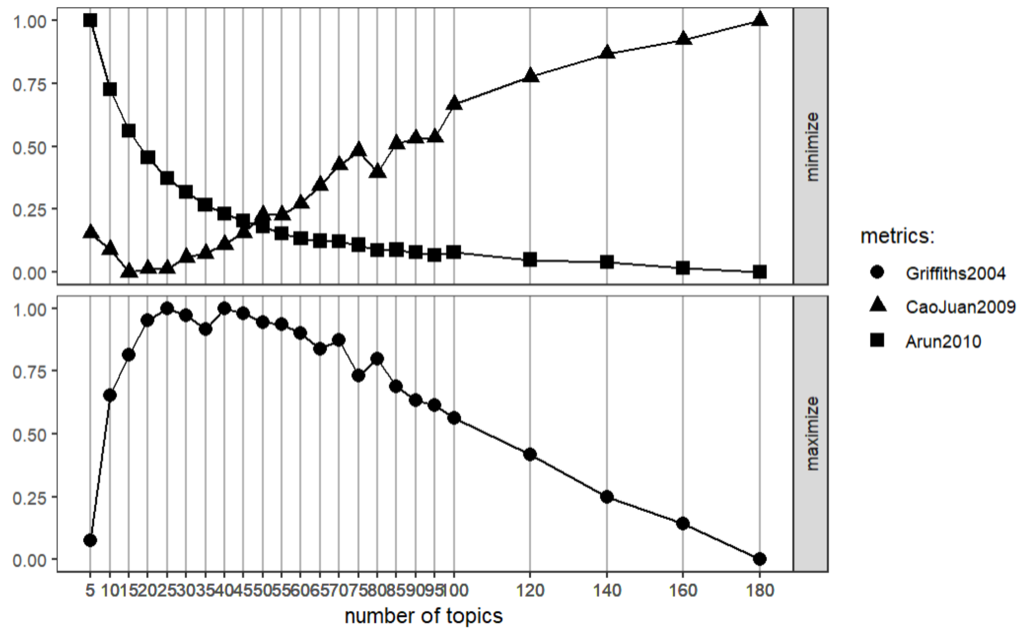

Fig 4. *Plot of the metrics of the LDA model for determining the optimal number of topics.*

A grid search was performed to find the optimal value of the parameters α and β, with the assumption that the optimal number of topics is 25. The values in Table 3 represent the average coherence scores per parameter combination, calculated as the mean of ten runs of the LDA model. The coherence scores were in the range of -1067.47 and -1072.19. The highest value in this grid search is for α = 0.05 and β = 0.01.

Table 3
Mean coherence scores of different parameters α and β of the LDA model on the top 20 words.

	α				
β	0.01	0.05	0.1	0.5	2
0.01	-1072.19	-1067.47	-1071.76	-1068.05	-1069.53
0.05	-1071.31	-1067.83	-1071.94	-1071.54	-1069.53
0.1	-1069.98	-1070.60	-1068.57	-1070.52	-1070.27
0.5	-1071.01	-1069.19	-1071.20	-1069.53	-1070.29

Figure 5 shows a plot of the calculated entropy scores of the BTM model for *k* topics in a range of 15 to 35. The graph shows a drop to 15 topics.

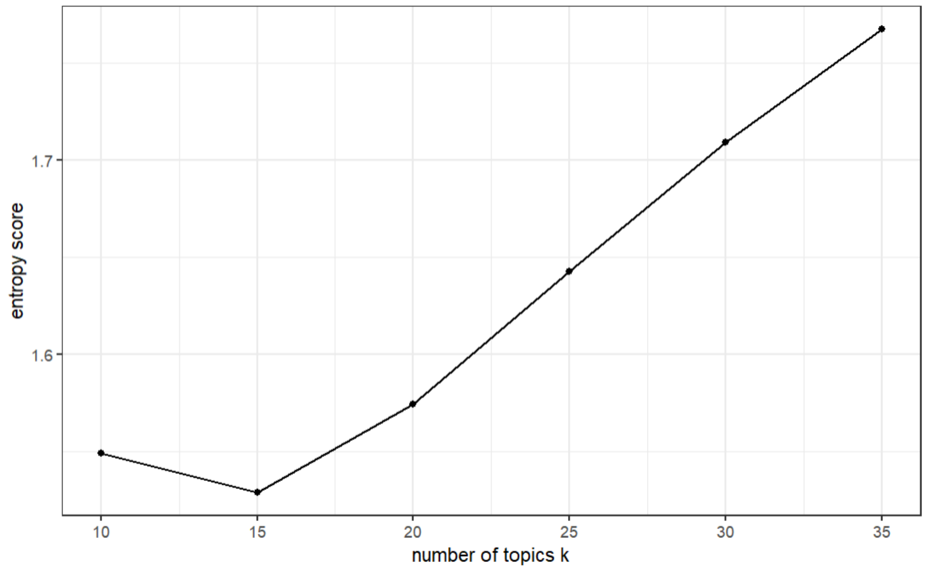

Fig 5. *Plot of the mean entropy score of the numbers of topics k.*

A grid search was performed for the BTM model as well, to identify the optimal values of parameters α and β. Assumed is an optimal number of topics of 15. Table 4 displays the average coherence scores per parameter combination, computed as the mean of five runs of the BTM model. The coherence scores ranged from -538.60 to -514.31. The highest score in this grid search was observed for α = 0.1 and β = 0.01.

Table 4
Mean coherence scores of different parameters α and β of the BTM model on the top 20 words.

	α				
β	0.01	0.05	0.1	0.5	3.3
0.01	-525.05	-523.35	-514.31	-523.46	-527.11
0.05	-521.57	-521.04	-518.97	-523.82	-518.18
0.1	-524.75	-529.86	-525.72	-519.91	-518.71
0.5	-537.40	-533.88	-538.60	-518.59	-528.88

5.2 Evaluation of the models

To evaluate the LDA model (k = 25, α = 0.05 and β = 0.01) and the BTM (k = 15, α = 0.1 and β = 0.01), the mean coherence scores were computed after running the models. Table 5 presents these scores, with three different values of the number of top topic words used for computing this. The mean coherence scores, along with the corresponding 95% confidence intervals (95%CI), are reported for each model. In terms of coherence, a higher mean coherence score indicates better performance. Comparing the two models, it could be observed that BTM consistently outperforms LDA across all values of T.

Table 5

Average coherence score on the top T words in topics discovered by LDA and BTM.

T	5		10		20	
	Mean	95%CI	Mean	95%CI	Mean	95%CI
LDA	-36.57	[-43.29, -30.54]	-217.58	[-243.13, -177.31]	-1076.15	[-1172.84, -964.66]
BTM	-23.07	[-28.02, -18.90]	-113.32	[-133.92, -99.76]	-514.31	[-573.51, -475.00]

Table 6 shows the top ten topic words of the LDA model. An employee of Scorius was asked to label each topic according to those ten words. Meaningful labels were assigned to only 12 out of the 25 topics. Only 4 of those topics exhibited clear themes or concepts that could be easily understood and interpreted, according to this employee. The other topics were found to be less interpretable, making it challenging to assign coherent labels. Also, the labeled topics had some overlap, as “communicatie” (communication) was labeled twice.

In the analysis of the top 10 topic words for each generated topic, notable observations were made. There is some overlap in the top ten words across the topics, for example the word “betrekken” (involve) could be seen in more than one topic. Also, the word “client” had four different versions in topic 11 and 16. Moreover, a significant number of words within the top 10 topic words were general adjectives and verbs. Examples of such words include "goed" (good), "gaan" (go), and "blijven" (stay).

Table 6

List of topics identified by the LDA model and manually labeled.

Topic id	Top 10 topic words	Labels given
1	elkaar kennis samenwerking gedrag leren zorgen ervaring niveau personen voorbeeld	Teamwork
2	mee bezig gedaan denk medewerker manager allemaal teamleider denken krijgen	-
3	organisatie leidinggevende waardoor plaats zeer horen zowel inzet moment brengen	-
4	personeel zorgen werkdruk goed groep houden komen groepen plannen eerlijk	-
5	werken werknemers dag werkt kantoor week klanten dagen jaar belangrijk	-
6	werk krijg werkzaamheden blijven mensen komt problemen tevreden werken mee	-
7	moment visie strategie processen nodig organisatie keuzes bijvoorbeeld doelen	Visie
8	goed gaat staat voldoende betrekken gewerkt gebeuren gedaan periode vele	-
9	communicatie betere duidelijk transparantie integratie openheid veranderingen kort bieden	Communicatie

10	collega's goede systeem management komt graag huidige inzetten belangrijk motivatie	-
11	tijd zorg cliënten taken uren extra krijgen geld cliënt groep	Zorg
12	betrokkenheid bedrijf betrokken vraag zeer groot werk vergroten voel betrekken	Betrokkenheid
13	nieuwe komen ontwikkelingen sneller proces huidige afdelingen systemen ICT diensten	Processen
14	mensen werkvloer luisteren managers staan gebeurt serieus volgen zorgen stap	-
15	zaken gevoel nodig vind snel zitten hierdoor veranderingen lang krijgen	-
16	middelen voldoende mondkapjes corona handschoenen collega beschermende bescherming clienten client	Corona
17	jaar salaris functie jaren blijft vaste dezelfde vast duidelijkheid contract	Arbeidsvoorwaarden
18	geven aandacht persoonlijke ontwikkeling ruimte mogelijkheden stellen duidelijk ontwikkelen komen	Persoonlijke ontwikkeling
19	beter communiceren blijven beleid daarnaast financieel opleiding transparant vind onderdeel	Communicatie
20	medewerkers waardering tonen leiding app verbeteren veiligheid bepaalde voelen gaan	-
21	vragen overleg contact uur afspraken gesprek mis huis goede familie	-
22	gaan laten zien directie laat staan vertrouwen vaker mee problemen	-
23	team teams duidelijkheid echt vind duidelijk gaat kijken komt terug	-
24	projecten werken juiste project grote afdeling manier uitvoering management	Projectmanagement
25	informatie delen weten vinden dingen info mail graag gebruik teveel	Mailcontact

Table 7 presents the top ten topic words of the labeled topics generated by the BTM model. The employee of Scorius labeled the topics again. Remarkably, all 15 topics were successfully labeled, with 8 topics considered a good fit. These 8 topics demonstrated clear and easily understandable themes according to the employee's assessment. The remaining topics were also labeled, but were relatively less interpretable, posing some challenges in assigning coherent labels.

In the analysis of the top 10 topic words, noteworthy observations emerged. Firstly, it was apparent that there was minimal overlap in the top 10 words across different topics. Secondly, the words consisted of a lot of nouns and meaningful adjectives such as "heldere" (clear) and "sociale" (social).

Table 7

List of topics identified by the BTM model and manually labeled.

Topic id	Top 10 topic words	Labels given
1	week dagen dag thuis telewerk verlof balans privé thuiswerk flexibiliteit	Werk-privé balans
2	klanten concurrentie enorme sociale schip houdt vorig wilt service regionale	Commercieel / klanten
3	functie opleidingen kwaliteiten beloning eerlijke kunde beoordeling hbo aanbod verdeling	Ontwikkeling
4	visie strategie koers stimuleren leiderschap heldere uitdragen concreet ondernemen doelstellingen	Leiderschap
5	motivatie schaal vergroot werkgever type beloond salarisverhoging cao rug pensioen	Arbeidsvoorwaarden
6	teamleider ervaar steun verhaal themamanager belangen grensoverschrijdend thema theorie grens	Veilige werkomgeving
7	aanspreken fouten kennen gezamenlijk expertise bijdrage creeren veilig managen vak	Teamwork
8	beslissingen gehoord bestuurder gevende onnodig hogerhand verlies luister bovenaf leeft	Management
9	corona route clienten ging vond kwam besmet ziek bang route	Corona

10	veranderingen onduidelijk wijzigingen overname kleding ontwerp eisen benodigde verwachtingen opgepakt	Veranderingen
11	nieuws app berichten kanalen innovaties algemene apps nieuwsbrieven social digitale	Communicatie
12	systemen uitvoering integraal inziens meerwaarde voortgang realiseren dienen projectteam primaire	Projectmanagement
13	mondkapjes middelen handschoenen beschermende bescherming handgel voorraad kregen test beschermend	Beschermings-middelen
14	uur casus psychiatrie telefonisch contacten declareren patiënt dementie minuten telefonische	Zorgcontact
15	taken teams groepen zelfsturend zelfsturende koste administratieve gezond begeleiders coach	Zelfsturende teams

6. Discussion and conclusion

The aim of this research is to answer the following question: “Are the topic modeling methods LDA and BTM effective to use to gain insight in the topics present in unstructured answers to open ended questions in employee engagement surveys?” In order to answer this question, the parameters has to be optimized and the models had to be evaluated first.

The optimization of topic modeling parameters is crucial in achieving meaningful results. In this study, an optimal number of 25 topics was determined for the LDA model, while the BTM model demonstrated the lowest entropy score for 15 topics. This divergence is expected considering the different approaches employed by the models.

During the parameter optimization process, an optimal value of 0.01 is identified for β, indicating a reduced emphasis on including all words within each topic. Such a small β value aligns well with our dataset, as it is expected that there are specific identifiable topics related to employee engagement. The value of α, which controls the topic distribution within documents, shows varying optimal values across the models, yet all fell within the moderate range. These values prioritize the presence of more dominant topics within each document. This suits our dataset, as individuals might express their opinions on a single subject or provide feedback on a mixture of topics in their answers.

The results of the topic modeling methods LDA and BTM with their top 10 corresponding words are presented in the results. According to the coherence scores, the BTM model is obviously more accurate in creating coherent topics. In the topic labeling assignment can be seen that the BTM was better interpretable and created understandable and coherent topics. Furthermore, the LDA model shows repetition of words across multiple topics, which indicates a lack of distinctiveness and specificity in the top word distributions. The top words also include a lot of general words that do not provide specific information or context that would aid in understanding the underlying topic. On the other hand, the nouns and adjectives in the BTM model indicate a higher level of distinctiveness and specificity in the word distributions.

To sum up, all three measures indicate that the performance of the BTM model was better than the LDA model. Nevertheless, it is important to note that there are some remarks on the BTM model as well. While 7 out of the 15 labels are less interpretable, and some top words appeared odd, indicating the need for manual adjusting before utilizing these topics in practical applications. To answer the research question, the findings highlight the superiority of the BTM model over the LDA model to create distinct and coherent topics in our specific research context. The BTM model is effective in gaining insights in the latent topics in unstructured answers to open ended questions in employee engagement surveys.

Despite the valuable insights gained from this research, it is important to acknowledge certain limitations. Firstly, it should be noted that different preprocessing techniques may yield different results. For instance, incorporating stemming in the preprocessing step could potentially avoid for example the issue of word repetition observed in the LDA model. By reducing words to their root forms, stemming may enhance the distinctiveness and specificity of the word distributions, leading to improved topic coherence. Future research could investigate the effectiveness of other preprocessing techniques.

Another limitation lies in the subjectivity involved in two out of three measures used for evaluation, which heavily rely on human judgment. Human interpretation of topics and coherence scores can introduce biases and variations, potentially influencing the assessment of model performance.

Additionally, this research focused on utilizing two well-established topic modeling methods, namely LDA and BTM. However, numerous other topic modeling models do exist, including state-of-the-art methods and variations of these two models, that might offer superior topic extraction capabilities. Exploring alternative models could potentially yield better results. Future research could explore the application of alternative models.

In order to get more information out of this text data, future research could integrate sentiment analysis and text classification techniques. By incorporating sentiment analysis, topics could be linked to positive or negative attitudes. Additionally, text classification methods could assist in categorizing documents into predefined topic categories, further enhancing the interpretability and applicability of topic modeling results.

6.1 Related work

In the research of Nanda (2021), LDA served as a way to discover latent topics in survey data. However, the study acknowledges the possibility of other relevant underlying themes that may exist. By solely focusing on one topic modeling method, the potential for other models to better fit the data is not explored. Consequently, our study takes advantage of comparing multiple methods. The comparison of more methods, as in our study, thereby increasing the likelihood of identifying a more comprehensive set of topics.

In the research conducted by Vidal, Ares, and Jaeger (2022), the BTM model successfully revealed labeled topics in responses to open-ended questions. Comparable results could be found in the assessment of the effectiveness of the BTM model in our study. However, it should be noted that the optimization of the parameters α and β was not conducted in the previously mentioned study. By optimizing these parameters, a better performance could be achieved.

Yan et al. (2013) outlined that the performance of the BTM model is superior to LDA, particularly when applied to short texts. Our study demonstrates comparable results to previous research while incorporating a combination of short and long texts. This finding provides valuable insights into the wider applicability of the BTM model beyond just short texts.

Overall, by considering the limitations of focusing solely on one method and incorporating optimization techniques, our research builds upon prior studies and offers a more comprehensive perspective on topic modeling.

6.2 Ethical implications and consideration

When conducting research, it is crucial to be aware of the ethical issues. Access to the dataset was given after signing a data security agreement, which is a contract drawn up by the company Scorius. By signing this contract, it is ensured that the handling of confidential information is done securely and that appropriate security measures are implemented to store the data. Furthermore, the data used for this research contains opinions and feelings of real individuals. This is sensitive information, and it is important that it cannot be traced back to any specific individual. All identifying information was excluded from this dataset and sensitive information in the text fragments was anonymized. In this way the privacy of the individuals was safeguarded.

6.3 Data science problem

The BTM model is used to create a topic list with the labeled topics and the twenty corresponding top words. The BTM model has served as a solid base for creating a topic list with corresponding topic words, but manual finetuning is needed to get a coherent and human interpretable list. 16 topics are included in the list with some manual adjustments in adding, renaming and deleting topics and words (see appendix C). This sets the stage to use the identified topics in analyzing the open-ended answers, for example with topic classification and sentiment analysis techniques. Scorius can utilize this list as a foundation for its analysis. The list can be continuously expanded by incorporating additional words and by adding translations, answers in other languages can also be analyzed. This ongoing enhancement of the list enables Scorius to effectively analyze a wider range of questions and responses.

6.4 Conclusion

One of the key challenges in analyzing open-ended answers is the labor-intensive nature, which typically requires significant effort. To address this challenge, this research investigated the potential of NLP techniques, specifically topic modeling, to automate the discovery of topics in unstructured text answers. The study explored two different topic modeling methods, namely LDA and BTM, to assess their effectiveness in uncovering latent topics. By employing these methods, the research aimed to automate the extraction of meaningful themes from the open-ended responses. The data underwent preprocessing, and the models were fine-tuned with optimized parameters. The LDA model failed to provide meaningful insights into the underlying topics. However, the results obtained from the BTM model proved to be highly valuable in extracting latent topics from unstructured and unlabeled text data. The BTM model, which employs biterms to address sparse word co-occurrence in short texts, successfully generated topics with interpretable sets of top words. With some manual adjustments and labeling of these topics, the outcomes can be effectively applied in the analysis of open-ended responses, for example when combined with topic classification and sentiment analysis techniques. This research contributes to the field of topic modeling by highlighting the effectiveness of the BTM model in analyzing unstructured text data from employee engagement surveys. The BTM model is a promising model to uncover topics in data with different text lengths and structures.

References

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I 14* (pp. 391-402). Springer Berlin Heidelberg

Benoit K, Muhr D, Watanabe K (2021). stopwords: Multilingual Stopword Lists_. R package version 2.3, https://CRAN.R-project.org/package=stopwords.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774-774.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Bystrov, V., Naboka, V., Staszewska-Bystrova, A., & Winker, P. (2022). Choosing the Number of Topics in LDA Models--A Monte Carlo Comparison of Selection Criteria. *arXiv preprint arXiv:2212.14074*.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775-1781.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2928-2941.

Fielding, J., Fielding, N., & Hughes, G. (2013). Opening up open-ended survey data using qualitative software. *Quality & Quantity*, *47*, 3261-3276.

Findley Musgrove, C., E. Ellinger, A., & D. Ellinger, A. (2014). Examining the influence of strategic profit emphases on employee engagement and service climate. *Journal of workplace learning*, *26*(3/4), 152-171.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl_1), 5228-5235.

Grün, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30, https://doi.org/10.18637/jss.v040.i13

Kaur, J., & Buttar, P. K. (2018). A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, *4*(4), 207-210.

Koltcov, S. (2018). Application of Rényi and Tsallis entropies to topic modeling optimization. *Physica A: Statistical Mechanics and its Applications*, *512*, 1192-1204.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*, 15169-15211.

Markos, S., & Sridevi, M. S. (2010). Employee engagement: The key to improving performance. *International journal of business and management*, *5*(12), 89.

Miles, M. B. (1979). Qualitative data as an attractive nuisance: The problem of analysis. *Administrative science quarterly*, *24*(4), 590-601.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).

Nanda, G., Douglas, K. A., Waller, D. R., Merzdorf, H. E., & Goldwasser, D. (2021). Analyzing large collections of open-ended feedback from MOOC learners using LDA topic modeling and qualitative analysis. *IEEE Transactions on Learning Technologies*, *14*(2), 146-160.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 100-108).

Nikita, M. (2020). ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. R package version 1.0.2, https://CRAN.R-project.org/package=ldatuning.

Ooms J (2022). cld2: Google's Compact Language Detector 2. R package version 1.2.4, https://CRAN.R-project.org/package=cld2.

Python Software Foundation (2023). Python Language Reference, https://www.python.org

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Schofield, A., & Mimno, D. (2016). Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, *4*, 287-300.

Terpilowski, M. (2021). Bitermplus. Python package version 0.7.0, https://bitermplus.readthedocs.io/en/latest/tutorial.html.

Vidal, L., Ares, G., & Jaeger, S. R. (2022). Biterm topic modelling of responses to open-ended questions: A study with US consumers about vertical farming. Food Quality and Preference, 100, 104611.

Welch, M. (2011). The evolution of the employee engagement concept: communication implications. *Corporate Communications: An International Journal*, *16*(4), 328-346.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445-1456).

Appendix A. Engagement monitor questions

The dataset includes the following questions:

1. Wat kan *Bedrijf* doen om jouw betrokkenheid te vergroten?
2. Wat kan het MT van *Bedrijf* *bedrijf* doen om de samenwerking binnen heel *Bedrijf* *bedrijf* te verbeteren?
3. Zijn er andere communicatiekanalen en -methoden waarvan je gebruik zou willen maken? Geef in het vak hieronder aan wat je ideeën daarover zijn
4. Zijn er andere thema's en onderwerpen waarover je graag geïnformeerd zou willen worden? Noteer je ideeën en suggesties in het vak hieronder
5. Zijn er andere communicatiekanalen en -methoden waarvan je gebruik zou willen maken? Geef in het vak hieronder aan wat je ideeën daarover zijn
6. Zijn er andere thema's en onderwerpen waarover je graag geïnformeerd zou willen worden? Noteer je ideeën en suggesties in het vak hieronder
7. Wat kan *bedrijf* doen zodat jij je werk beter kan doen?
8. Heb je nog andere feedback die je met ons wilt delen?
9. Wat kan de organisatie doen om jouw tevredenheid te verhogen?
10. Wat is er nodig om de integratie van **bedrijf** met **bedrijf** te doen slagen?
11. Is de huidige tarievenlijst wat jou betreft compleet? Mis je verrichtingen/activiteiten die je wel doet maar in het huidige systeem niet kwijt kunt? Zo ja: welke verrichtingen zijn dat dan en kun je een indicatie geven van hoeveel tijd je daar kwijt aan bent?
12. Wat gaat er goed binnen **bedrijf**?
13. Wat draag jij bij aan **bedrijf**?
14. Wat vind jij dat er verbeterd kan worden binnen **bedrijf**?
15. Kun je een praktijkvoorbeeld geven van een werksituatie waarin collega's elkaar niet respectvol hebben behandeld binnen **bedrijf**?
16. Wil je nog iets kwijt wat niet in de vragenlijst aan de orde is gekomen?
17. Kun je ons vertellen waarom je dat zo voelt?
18. Is er nog iets dat je wilt delen, m.b.t. de transformatie naar een High Performance Culture?
19. Wil je nog iets aan ons kwijt, als het gaat over interne communicatie?
20. Is er nog iets dat je wilt delen, m.b.t. de transformatie naar een meer Adaptive Culture?
21. Kun je ons vertellen waarom je dat vind?
22. Is er iets dat je zou willen delen over de transformatie richting een meer adaptieve cultuur dat we je niet hebben gevraagd?
23. Zijn er eventueel andere zaken die u in deze enquête graag zou willen aanhalen?
24. Wat had er beter gekund tijdens de corona periode?
25. Wat ging er goed tijdens de Corona periode?
26. Wat ga je als eerste uit de Vitaal vakmanschap vitaliteitstraining in de praktijk brengen?
27. Wat kan **bedrijf** doen om je nog meer te ondersteunen bij het thema vitaliteit?
28. Is er iets in je werk waar je niet tevreden over bent? En kan je vertellen waarom niet?
29. Wat moet **bedrijf** meer doen om grensoverschrijdend gedrag te voorkomen?

30. Heb je een opmerking of wil je een toelichting geven?
31. Heb je opmerkingen, tips en/of aanvullingen? Geef dit dan hieronder weer:
32. Wat hadden we beter kunnen doen tijdens de Coronaperiode?

Appendix B. Used software

Most of the analyses were performed using R Statistical Software (v4.2.1; R Core Team 2023). The following packages were used:

- Filtering on Dutch language: cld2 package (v1.2.4; Ooms 2022)
- Tokenization: quanteda package (v3.3.1; Benoit et al. 2018)
- Stopwords: stopwords package (v2.3; Benoit et al. 2021)
- LDA model: topicmodels package (v0.2-14, Grün & Hornik 2023)

The analysis of the BTM model was performed with Python (v3.10.12; Python Software Foundation 2023). Python is used in the online coding environment Google Colab. The used package is bitermplus (v0.7.0; Terpilowski 2021).

The code used for this research can be downloaded at Github.
[CarlijnD/ADSThesisLDABTM (github.com)](https://github.com/CarlijnD/ADSThesisLDABTM)

Appendix C. Topic list

The BTM model served as base for creating a topic list. 16 topics are included in the list with some manual adjustments in adding and deleting topics and words.

Topics	Words
Werk-privé balans	week dag thuis telewerk verlof balans privé thuiswerk flexibiliteit werkuren vrije reiskosten woning overuren gezin kinderen opvang flexibel
Opleiding & Ontwikkeling	functie opleidingen kwaliteiten hbo aanbod doorgroei anciënniteit doorgroeimogelijkheden doorgroeien wensen wens mbo universiteit training cursus coaching
Leiderschap	visie strategie koers leiderschap heldere concreet ondernemen doelstellingen competenties kernwaarden directies transparant ondernemerschap
Waardering werk	motivatie ingehuurd werkgever schouderklopje vergroot afscheid jubileum gewaardeerd waardering compliment gezien gehoord
Psychologische veiligheid	steun belangen grensoverschrijdend grens respectvol oor zware verhaal geluisterd luisterend pesten intimidatie intimideren discriminatie uitschelden respectloos respect ongepast toezicht vertrouwenspersoon grensoverschrijdend
Teamwork	aanspreken fouten kennen gezamenlijk expertise bijdrage managen vak kritische kwalitatief collegialiteit respecteren knelpunten sessies benaderbaar nadruk projecten samenwerken teamwork team teamleider collega's taken teams groepen zelfsturend zelfsturende administratieve begeleiders coach begeleiden zelforganiserende coaches
Management	beslissingen gehoord bestuurder onnodig hogerhand luister bovenaf leeft mist regie opgelegd voorstellen deskundige wandelgangen oprecht manager leidinggevende leiding instructies
Corona	corona route clienten besmet ziek bang routes getest covid beschermingsmiddelen kwetsbare ouderen
Veranderingen	veranderingen onduidelijk wijzigingen overname benodigde verwachtingen doorgevoerde innovatie communicatie fusie samenvoegen implementatie
Communicatie	nieuws app berichten kanalen innovaties apps nieuwsbrieven social digitale applicaties linkedin buitenland media communicatie communiceren online intranet platform post facebook instagram twitter
Project-management	systemen uitvoering integraal meerwaarde voortgang realiseren projectteam controle effectief beheer projecten project lean management projectmatig projectleider projectmanager
Beschermings-middelen	mondkapjes middelen handschoenen beschermende bescherming handgel voorraad test beschermd mondkapjes beschermen instructie handalcohol beschikking

Zorgcontact	casus psychiatrie telefonisch contacten declareren patiënt dementie telefonische patient verslaglegging psychiatrische casuistiek telefoontjes overlijden
Faciliteiten	auto kantine laptop telefoon internet schoonmaakmiddelen bureau werkplekken parkeren kantoor middelen bescherming fietsplan gezond
Rooster	dag week uren flexibel schema planning agenda onregelmatig tijden
Arbeids-voorwaarden	salaris schaal beloning cao voorwaarden pensioen geld loon verhoging onderhandeling